

Learning the Gain Values and Discount Factors of DCG

Ke Zhou
Dept. of Computer Science
and Engineering
Shanghai Jiao-Tong University
No. 800, Dongchuan Road,
Shanghai, China 200240
zhouke
@apex.sjtu.edu.cn

Hongyuan Zha
College of Computing
Georgia Institute of
Technology
Atlanta, GA 30032
zha@cc.gatech.edu

Gui-Rong Xue, Yong Yu
Dept. of Computer Science
and Engineering
Shanghai Jiao-Tong University
No. 800, Dongchuan Road,
Shanghai, China 200240
{grxue,
yyu}@apex.sjtu.edu.cn

ABSTRACT

Evaluation metrics are an essential part of a ranking system, and in the past many evaluation metrics have been proposed in information retrieval and Web search. Discounted Cumulated Gains (DCG) has emerged as one of the evaluation metrics widely adopted for evaluating the performance of ranking functions used in Web search. However, the two sets of parameters, gain values and discount factors, used in DCG are determined in a rather ad-hoc way. In this paper we first show that DCG is generally not coherent, meaning that comparing the performance of ranking functions using DCG very much depends on the particular gain values and discount factors used. We then propose a novel methodology that can learn the gain values and discount factors from user preferences over rankings. Numerical simulations illustrate the effectiveness of our proposed methods. Please contact the authors for the full version of this work.

1. INTRODUCTION

Discounted Cumulated Gains (DCG) is a popular evaluation metric for comparing the performance of ranking functions [4]. It can deal with multi-grade judgments and it also explicitly incorporates the position information of the documents in the result sets through the use of discount factors. However, in the past, the selection of the two sets of parameters, gain values and discount factors, used in DCG is rather arbitrary, and several different sets of values have been used. This is rather an unsatisfactory situation considering the popularity

of DCG. In this paper, we address the following two important issues of DCG:

1. Does the parameter set matter? I.e., do different parameter sets give rise to different preference over the ranking functions?
2. If the answer to the above question is *yes*, is there a principled way to selection the set of parameters?

The answer to the first question is *yes* if there are more than two grades used in the evaluation. This is generally the case for Web search where multiple grades are used to indicate the degree of relevance of documents with respect to a query. We then propose a principled approach for learning the set of parameters using preferences over different rankings of the documents. As will be shown the resulting optimization problem for the learning the parameters can be solved using quadratic programming very much like what is done in support vector machines for classification. We did several numerical simulations that illustrate the feasibility and effectiveness of the proposed methodology. We want to emphasize that the experimental results are preliminary and limited in its scope because of the use of the simulation data; and experiments using real-world search engine data are being considered.

2. RELATED WORK

Cumulated gain based measures such as DCG [4] have been applied to evaluate information retrieval systems. Despite their popularity, little research has been focused on analyzing the coherence of these measures to the best of our knowledge. The study of [7] shows that different gain values of DCG can raise different judgements of ranking lists. In this study, we first prove that the DCG is incoherency and then propose a principled method to learn the DCG parameters as a linear utility function.

Learning to rank attracts a lot of research interests in recent years. Several methods have been developed to learn the ranking function through directly optimization performance metrics such as MAP and DCG [5, 8, 9].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

These studies focus on learning a good ranking function with respect to given performance metrics, while the goal of this paper is to analysis coherence of DCG and propose a learning method to determine the parameters of DCG.

As we have mentioned in Section 5, DCG can be viewed as a linear utility function. Therefore, the problem of learning DCG is closely related to the problem of learning the utility function. Learning utility function is studied under the name of conjoint analysis by the market science community [1, 6]. The goal of conjoint analysis is to model the users' preference over products and infer the features that satisfy the demands of users. Several methods have been proposed to model solve the problem [2, 3].

3. DISCOUNTED CUMULATED GAINS

We first introduce some notation used in this paper. We are interested in ranking N documents $\mathcal{X} = \{x_1, \dots, x_N\}$. We assume that we have a finite ordinal label (grade) set $\mathcal{L} = \{\ell_1, \dots, \ell_L\}$. We assume that ℓ_i is preferred over ℓ_{i+1} , $i = 1, \dots, L-1$. In Web search, for example, we can have

$$\mathcal{L} = \{\text{Perfect, Excellent, Good, Fair, Bad}\},$$

i.e., $L = 5$. A ranking of \mathcal{X} is a permutation

$$\pi = (\pi(1), \dots, \pi(N)),$$

of $(1, \dots, N)$, i.e., the rank of $x_{\pi(i)}$ under the ranking π is i .

For each label is associated a *gain value* $g_i \equiv g(\ell_i)$, and $g_i, i = 1, \dots, L$ constitute the set of gain values associated with \mathcal{L} . The DCG for π with the associated labels is computed as

$$DCG_{g,K}(\pi) = \sum_{i=1}^K c_i g_{\pi(i)}, \quad K = 1, \dots, N,$$

where $c_1 > c_2 > \dots > c_K > 0$ are the so-called *discount factors* [4].

The gain vector $g = [g_1, \dots, g_L]$ is said to be *compatible* if $g_1 > g_2 > \dots > g_L$. If two gain vectors g^A and g^B are both compatible, then we say they are compatible with each other. In this case, there is a transformation ϕ such that

$$\phi(g_i^A) = g_i^B, \quad i = 1, \dots, L,$$

and the transformation ϕ is *order preserving*, i.e.,

$$\phi(g_i) > \phi(g_j), \quad \text{if } g_i > g_j.$$

4. INCOHERENCY OF DCG

Now assume there are two rankers A and B using DCG with gain vectors g^A and g^B , respectively. We want to investigate how coherent A and B are in evaluating different rankings

4.1 Good News

First the good news: if A and B are compatible, then A and B agree on which set of rankings is optimal, i.e., which set of rankings have the highest DCG. We first state the following well-known result.

Proposition 1. Let $a_1 \geq \dots \geq a_N$ and $b_1 \geq \dots \geq b_N$. Then

$$\sum_{i=1}^N a_i b_i = \max_{\pi} \sum_{i=1}^N a_i b_{\pi(i)}.$$

It follows from the above result that any ranking π such that

$$g_{\pi(1)} \geq g_{\pi(2)} \geq \dots \geq g_{\pi(K)}$$

achieves the highest $DCG_{g,K}$, as long as the gain vector g is compatible.

How about those rankings that have smaller DCGs? We say two compatible rankers A and B are *coherent*, if they score any two rankings coherently, i.e., for rankings π_1 and π_2 ,

$$DCG_{g^A,K}(\pi_1) \geq DCG_{g^A,K}(\pi_2)$$

if and only if

$$DCG_{g^B,K}(\pi_1) \geq DCG_{g^B,K}(\pi_2),$$

i.e., ranker A thinks π_1 is better than π_2 if and only if ranker B thinks π_1 is better than π_2 . Now the question is whether compatibility implies coherency. We have the following result.

Theorem. If $L = 2$, then compatibility implies coherency.

Proof. Fix $K > 1$, and let

$$c = \sum_{i=1}^K c_i.$$

When there are only two labels, let the corresponding gains be g_1, g_2 . For a ranking π , define

$$c_1(\pi) = \sum_{\pi(i)=\ell_1} c_i, \quad c_2(\pi) = c - c_1(\pi).$$

Then

$$DCG_{g,K}(\pi) = c_1(\pi)g_1 + c_2(\pi)g_2.$$

For any two rankings π_1 and π_2 ,

$$DCG_{g^A,K}(\pi_1) \geq DCG_{g^A,K}(\pi_2)$$

implies that

$$c_1(\pi_1)g_1^A + c_2(\pi_1)g_2^A > c_1(\pi_2)g_1^A + c_2(\pi_2)g_2^A$$

which gives

$$(c_1(\pi_1) - c_1(\pi_2))(g_1^A - g_2^A) > 0.$$

Since A and B are compatible, the above implies that

$$(c_1(\pi_1) - c_1(\pi_2))(g_1^B - g_2^B) > 0.$$

Therefore $DCG_{g^B,K}(\pi_1) \geq DCG_{g^B,K}(\pi_2)$. The proof is completed by exchange A and B in the above arguments.

4.2 Bad News

Not too surprisingly, compatibility does not imply coherency when $L > 2$. We now present an example.

Example. Let $\mathcal{X} = \{x_1, x_2, x_3\}$, i.e., $N = 3$. We consider $DCG_{g,K}$ with $K = 2$. Assume the labels of x_1, x_2, x_3 are ℓ_2, ℓ_1, ℓ_3 , and for ranker A , the corresponding gains are $2, 3, 1/2$. The optimal ranking is $(2, 1, 3)$. Consider the following two rankings,

$$\pi_1 = (1, 3, 2), \quad \pi_2 = (3, 2, 1)$$

None of them is optimal. Let the discount factors be

$$c_1 = 1 + \epsilon, \quad c_2 = 1 - \epsilon, \quad 1/4 < \epsilon < 1.$$

It is easy to check that

$$\begin{aligned} DCG_{g^A,2}(\pi_1) &= 2c_1 + (1/2)c_2 > \\ &> (1/2)c_1 + 3c_2 = DCG_{g^A,2}(\pi_2). \end{aligned}$$

Now let $g^B = \phi(g^A)$, where $\phi(t) = t^k$, and ϕ is certainly order preserving, i.e., A and B are compatible. However, it is easy to see that for k large enough, we have

$$2^k c_1 + (1/2)^k c_2 < (1/2)^k c_1 + 3^k c_2$$

which is the same as

$$DCG_{g^B,2}(\pi_1) < DCG_{g^B,2}(\pi_2).$$

Therefore, A thinks π_1 is better than π_2 while B thinks π_2 is better than π_1 even though A and B are compatible. This implies A and B are *not* coherent.

4.3 Remarks

When we have more than two labels, which is the case for Web search, using DCG_K with $K > 1$ to compare the DCGs of various ranking functions will very much depend on the gain vectors used. Different gain vectors can lead to completely different conclusions about the performance of the ranking functions.

The current choice of gain vectors for Web search is rather *ad hoc*, and there is no criterion to judge which set of gain vectors are reasonable or natural.

5. LEARNING GAIN VALUES AND DISCOUNT FACTORS

DCG can be considered as a simple form of linear *utility function*. In this section, we discuss a method to learn the gain values and discount factors that constitute this utility function.

5.1 A Binary Representation

We consider a fixed K , and we use a binary vector $s(\pi)$ of dimension $K \times L$ to represent a ranking π considered for $DCG_{g,K}$. Here L is the number of levels

of the labels. In particular, the first L -components of s correspond to the first position of the K -position ranking in question, and the second L -components the second position, and so on. Within each L -components, the i -th component is 1 if and only if the item in position one has label $\ell_i, i = 1, \dots, L$.

EXAMPLE. In the Web search case, $L = 5$, suppose we consider $DCG_{g,3}$, and for a particular ranking π the labels of the first three documents are

Perfect, Bad, Good.

Then the corresponding 15-dimensional binary vector $s(\pi)$ is

$$[1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0].$$

We postulate a utility function $u(s) = w^T s$ which a linear function of s , and w is the weight vector, and we write

$$w = [w_{1,1}, \dots, w_{1,L}, w_{2,1}, \dots, w_{2,L}, \dots, w_{K,1}, \dots, w_{K,L}].$$

We distinguish two cases.

CASE 1. The gain values are position independent. This corresponds to the case

$$w_{i,j} = c_i g_j, \quad i = 1, \dots, K, j = 1, \dots, L.$$

This is to say that $c_i, i = 1, \dots, K$ are the discount factors, and $g_j, j = 1, \dots, L$ are the gain values. It is easy to see that

$$w^T s(\pi) = DCG_{g,K}(\pi).$$

CASE 2. In this framework, we can consider the more general case that the gain values are *position dependent*. Then $w_{1,1}, \dots, w_{1,L}$ are just the products of the discount factor c_1 and the position dependent gain values for position one, and so on. In this case, there is no need to separate the gain values and the discount factors. The weights in the weight vector w are what we need.

5.2 Learning w

We assume we have available a partial set of preferences over the set of all rankings. For example, we can present a pair of rankings π_1 and π_2 to a user, and the user prefers π_1 over π_2 , denoted by $\pi_1 \succ \pi_2$, which translates into $w^T s(\pi_1) \geq w^T s(\pi_2)$. Let the set of preferences be

$$\pi_i \succ \pi_j, \quad (i, j) \in S.$$

In the second case described above, we can formulate the problem as learning the weight vector w subject to a set of constraints (similar to rank SVM):

$$\min_{w, \xi_{ij}} w^T w + C \sum_{(i,j) \in S} \xi_{ij}^2 \quad (1)$$

subject to

$$w^T (s(\pi_i) - s(\pi_j)) \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0, \quad (i, j) \in S.$$

$$w_{kl} \geq w_{k,l+1}, \quad k = 1, \dots, K, \quad l = 1, \dots, L-1$$

For the first case, we can compute w as in Case 2, and then find c_i and g_j to fit w . It is also possible to carry out hypothesis testing to see if the gain values are position dependent or not.

6. SIMULATION

In this section, we report the results of numerical simulations to show the feasibility and effectiveness of the method proposed in Equation (1).

6.1 Experimental Settings

We use a ground-truth w to obtain preference of ranking lists. Our goal is to investigate whether we can reconstruct w via learning from the preference of ranking lists. The ground-truth w is generated according to the following equation:

$$w_{kl} = \frac{G_l}{\log(k+1)} \quad (2)$$

For a comprehensive comparison, we distinguish two settings of G_l . Specifically, we set $G_l = l$ in the first setting (Data 1), and $G_l = 2^l - 1$ in the second setting (Data 2).

The ranking lists are obtained by randomly permuting a ground-truth ranking list. For example, the ranking lists can be generated by permuting the list $[5, 5, 4, 4, 3, 3, 2, 2, 1, 1]$ randomly. We randomly generate different numbers of pairs of ranking lists and use the ground-truth to judge which ranking lists is preferred. Specifically, if $w^T s(\pi_1) > w^T s(\pi_2)$, we have a preference pair $\pi_1 \succ \pi_2$. Otherwise we have a preference pair $\pi_2 \succ \pi_1$.

6.2 Evaluation Measures

Given the estimated \hat{w} and the ground-truth w , we apply two measures to evaluate the quality of the estimated \hat{w} .

The first measure is the precision on a test set. A number of pairs of ranking lists are generated as the test set. We apply \hat{w} to predict the preference over the test set. Then the precision \hat{w} is calculated as the proportion of correctly predicted preference in the test set.

The second measure is the similarity of w and \hat{w} . Given the true value of w and the estimation \hat{w} defined by the above optimization problem, the similarity between w and \hat{w} can be defined as follows:

$$T(w) = (w_{11} - w_{1L}, \dots, w_{1L} - w_{1L}, \dots, w_{K1} - w_{KL}, \dots, w_{KL} - w_{KL}) \quad (3)$$

We can observe that the transformation T preserve the orders between ranking lists, i.e., $T(w)^T s(\pi_1) > T(w)^T s(\pi_2)$ iff $w^T s(\pi_1) > w^T s(\pi_2)$. The similarity between w and \hat{w} is measured by $\text{sim}(w, \hat{w}) = \frac{T(w)^T T(\hat{w})}{\|T(w)\| \|T(\hat{w})\|}$.

6.3 General Performance

We randomly sample a number of ranking lists and generate preference pairs according to a ground-truth w . The number of preference pairs in training set ranges from 20 to 200. We plot the precision and similarity of the estimated \hat{w} with respect to the number of training pairs in Figure 1 and Figure 2. It can be observed from Figure 1 and 2 that the performance generally grows with the increasing of training pairs, indicating that the preference over ranking lists can be utilized to enhance the estimation of the utility function w . Another observation is that when about 200 preference pairs are included in training set, the precisions in test sets become close to 95% under both settings. This observation suggests that we can estimate w precisely from the preference of ranking lists. We also notice that the similarity and precision sometimes give different conclusions of the relative performance over Data 1 and Data 2. We think it is because the similarity measure is sensitive to the choice of the offset constant. For example, large offset constants will give similarity very close to 1. Currently, we use w_{1L}, \dots, w_{KL} of the offset constant as in Equation (3). Generally, we refer the precision as a more meaningful evaluation metric and report similarity as a complement to precision.

After the utility function w is obtained, we can reconstruct the gain vector and the discount factors from w . To this end, we rewrite w as a matrix W of size $L \times K$. Assume that singular value decomposition of the matrix W can be expressed as $W = U \text{diag}(\sigma_1, \dots, \sigma_n) V^T$ where $\sigma_1 \geq \dots \geq \sigma_n$. Then, the rank-1 approximation of W is $\hat{W} = \sigma_1 u_1 v_1^T$. In this case, the first left singular vector u_1 is the estimation of the gain vector and the first right singular vector v_1 is the estimation of the discount factors. We plot the estimated gain vector and discount factors with respect to their true values in Figure 4 and Figure 3, respectively. Note the perfect estimations give straight lines in these two figures. We can see that the discount factors for the top-ranked positions are more close to a straight line and thus are estimated more accurately. This is because the discount factors of top-ranked positions have greater impact to the preference of the ranking lists. Therefore, these discount factors are captured more precisely by the constraints. The similar phenomenon can also be observed for the gain vector.

6.4 Noisy Settings

In real world scenarios, the preference pairs of ranking lists can be noisy. Therefore, it is interesting to investigate the effect of the noisy pairs to the performance. To this end, we fix the number training pairs to be 200 and create the noisy pairs by randomly flip a number of pairs in the training set. In our experiments, the number of noisy pairs ranges from 5 to 40. Since the trade off value C is important to the performance in the noisy setting, we select the value of C that shows

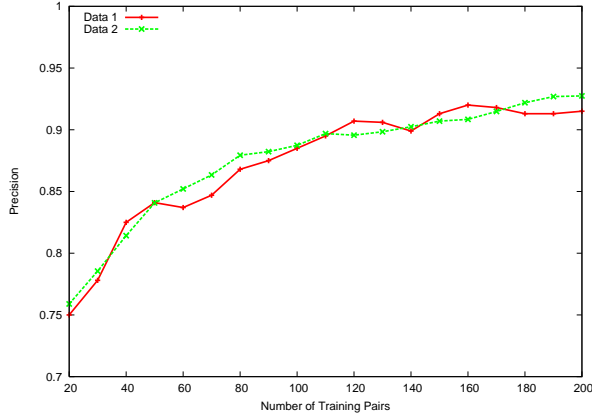


Figure 1: Precision over the test set with respect to the number of training pairs

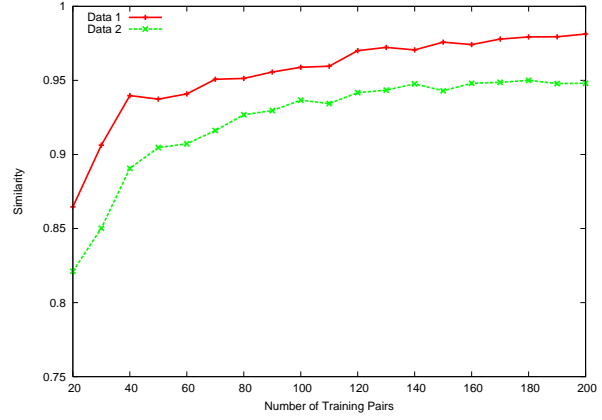


Figure 2: Similarity between w and \hat{w} with respect to the number of training pairs

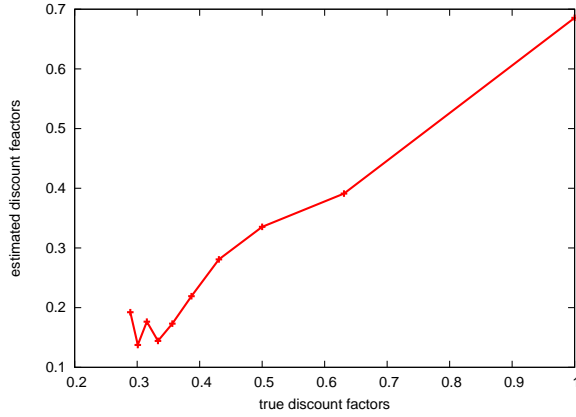


Figure 3: Estimated discount factors with respect to true discount factors

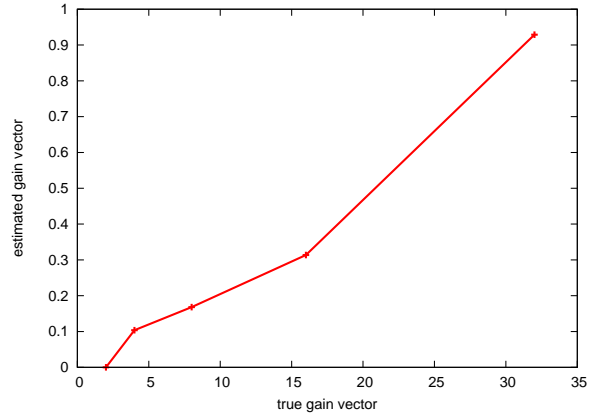


Figure 4: The estimated gain vector with respect to the true gain vector

the best performance on an independent validation set. We report the performance with respect to the number of noisy pairs in Figure 5 and Figure 6. We can observe that the performance decreases when the number of noisy pairs grows.

In addition to the noisy preference pairs, we also consider the noise in the grades of documents. In this case, we randomly modify the grades of a number of documents to form noise in training set. The estimated \hat{w} is used to predict to preference on a test set. The precision with respect to the number of noisy documents is shown in Figure 7. It can be observed that the performance decreases when the number of noisy documents grows.

6.5 Optimal Rankings

We can further restrict the preference pairs by involving the optimal ranking in each pair of training data. For example, we set one ranking list of the preference

pair to be the optimal ranking $[5, 5, 4, 4, 3, 3, 2, 2, 1, 1]$. In this case, if the other ranking list is generated by permuting the same list, it is implied by Proposition 1 that any compatible gain vectors will agree on the optimal ranking is preferred to other ranking lists. In other words, the preference pairs do not carry any constraints to the utility function w . Therefore, the constraints corresponding to these preference pairs are not effective in determining the utility function w . Consequently, the performance do not increase when the number of training pair grows as shown in Figure 9.

If the ranking lists contain different sets of grades, a fraction of constraints can be effective. The performance grows slowly with the number of training pairs increases as reported in Figure 10. By comparing Figure 1,2 and Figure 10, we can observe that when the type of preference is restricted, the learning algorithm requires more pairs to obtain a comparable performance.

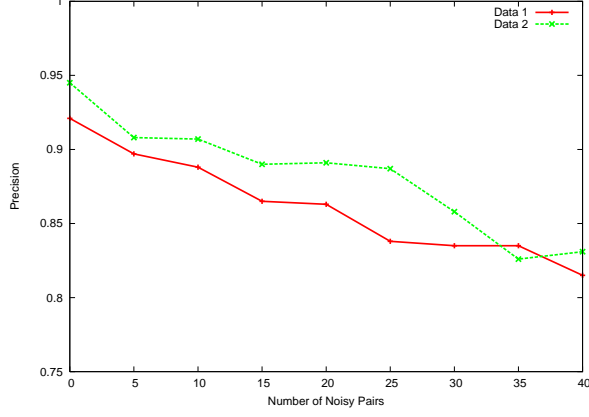


Figure 5: Precision over the test set with respect to the number of noisy pairs in noisy settings

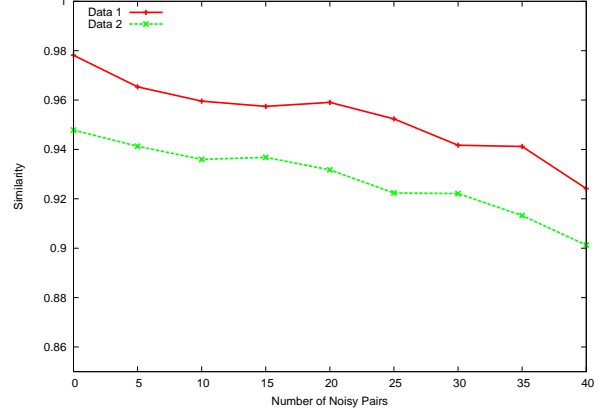


Figure 6: Similarity between w and \hat{w} with respect to the number of noisy pairs in noisy settings

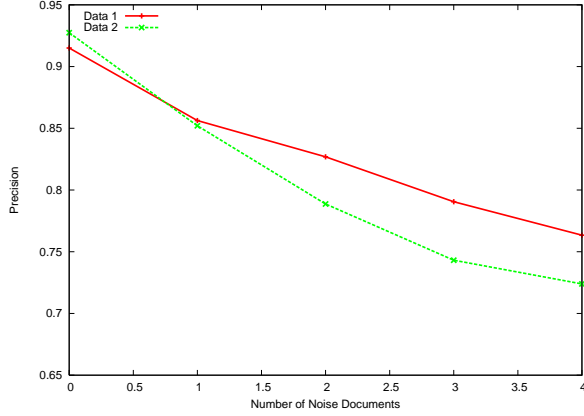


Figure 7: Precision with respect to the number of noisy grades in noisy settings

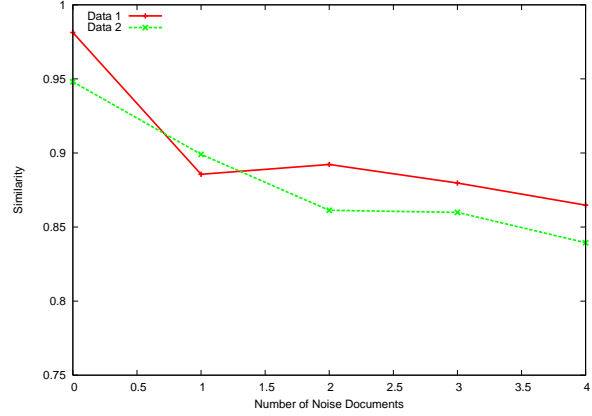


Figure 8: Similarity with respect to the number of noisy grades in noisy settings

We conclude from this observation that some pairs are more effective than others to determine w . Thus, if we can design algorithm to select these pairs, the number of pairs required for training can be greatly reduced. How to design algorithms to select effect preference pairs for learning DCG will be addressed as a future research topic.

7. AN ENHANCED MODEL

The objective function defined in Equation (1) does not consider the degree of difference between ranking lists. For example, it deals with preference pairs $[5, 5, 4, 2, 1] \succ [5, 4, 5, 2, 1]$ and $[5, 5, 4, 2, 1] \succ [2, 1, 5, 5, 4]$ in the same approach, although they have great differences in DCG. In order to overcome this problem, we propose an enhanced model that takes the degree of difference between ranking list into consideration.

$$\min w^T w + C \sum_{(i,j) \in S} \xi_{ij}^2 \quad (4)$$

subject to:

$$\begin{aligned} w^T (s(\pi_i) - s(\pi_j)) &\geq \text{Dist}(\pi_i, \pi_j) - \xi_{ij} \quad (i, j) \in S \\ \xi_{ij} &\geq 0 \quad (i, j) \in S \\ w_{kl} &\geq w_{k,l+1} \quad k = 1, \dots, K \text{ and } l = 1, \dots, L-1 \end{aligned}$$

where $\text{Dist}()$ is a distance measure for a pair of permutations π_1 and π_2 . In principle, we would prefer $\text{Dist}(\pi_i, \pi_j)$ to be a good approximation of $DCG(\pi_i) - DCG(\pi_j)$. However, since we do not actually know the ground-truth w in practice, it is generally difficult to obtain a precise approximation. In our simulation, we

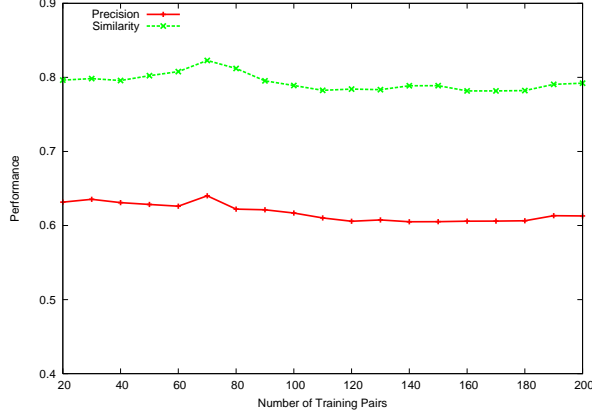


Figure 9: Performance when training pairs are generated by permuting the same list

apply the Hamming distance as the distance measure:

$$\text{Ham}(\pi_1, \pi_2) = \sum_k 1[\pi_1(k) \neq \pi_2(k)] \quad (5)$$

We perform simulation to evaluate the enhanced model. From Figure 12 and Figure 11, we can observe that the performance improvement of the enhanced model is not very significant. We doubt that this is because Hamming distance is not a precise approximation of the DCG difference. We plan to investigate this problem in our future study.

8. LEARNING WITHOUT DOCUMENT GRADES

When the grades of the documents are not available, we can also fit a model to predict the preference of two ranking list. To this end, we use a $K \times K$ -dimensional binary vector $s(\pi)$ to represents the ranking π . The first K components of $s(\pi)$ correspond to the first position of the K -position ranking, and the second K -components the second the position, and so on.

For example, for a ranking list π

$$d_3, d_1, d_4, d_2, d_5$$

the corresponding binary vector is

$$[0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1]$$

Given a set of preference over ranking lists, we can obtain w by solving the following optimization problem:

$$\min w^T w + C \sum_{(i,j) \in S} \xi_{ij}^2 \quad (6)$$

subject to:

$$w^T (s(\pi_i) - s(\pi_j)) \geq 1 - \xi_{ij} \quad (i, j) \in S \quad (7)$$

$$\xi_{ij} \geq 0 \quad (i, j) \in S \quad (8)$$

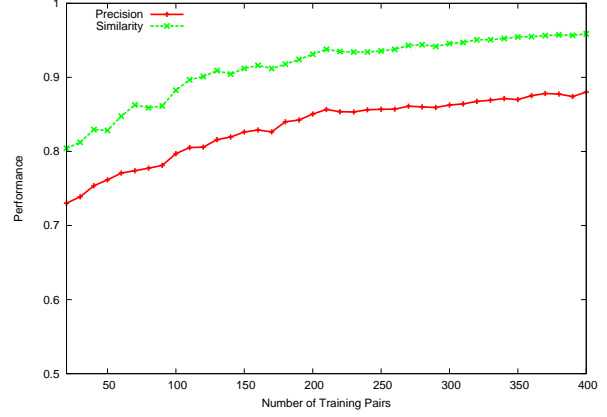


Figure 10: Performance when training pairs are generated from different lists

In this case, the constraints $w_{kl} \geq w_{k,l+1}$ are not included in the optimization problem, since we do not have any prior knowledge about the grades of the documents. The precision on test set with respect to the number of training pairs is reported in Figure 13. We can observe that the w can be precisely learned even without the grades of documents. In this case, the learned utility function w can be interpreted as the relevant judgements for the documents.

9. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate the coherence of DCG, which is an important performance measure in information retrieval. Our analysis show that the DCG is incoherency in general, i.e., different gain vectors can lead to different judgements about the performance of a ranking function. Therefore, it is a vital problem to select reasonable parameters for DCG in order to obtain meaningful comparisons of ranking functions. We propose to learn the DCG gain values and discount factors from preference judgements of ranking lists. In particular, we develop a model to learn DCG as a linear utility function and formulate the method as a quadratic programming problem. Preliminary results of simulation suggest the effectiveness of the proposed method.

We plan to further investigate the problem of learning DCG and apply the proposed method in real world data sets. Furthermore, we plan to generalize DCG to nonlinear utility functions to model more sophisticated requirements of ranking lists, such as diversity and personalization.

10. REFERENCES

- [1] J. D. Carroll and P. E. Green. Psychometric methods in marketing research: Part II, Multidimensional scaling. *Journal of Marketing Research*, 34:193–204, 1997.

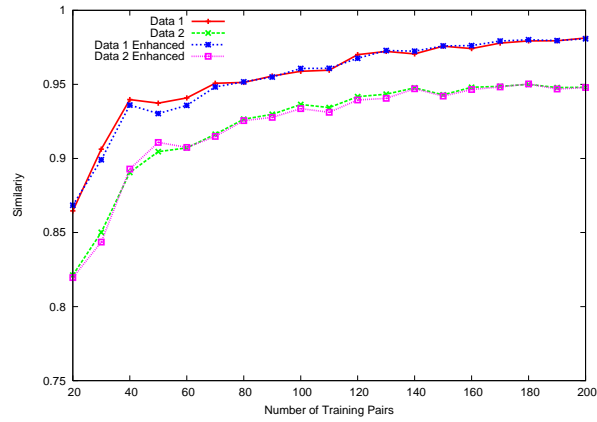


Figure 11: Similarity of the enhanced model with respect to the number of training pairs

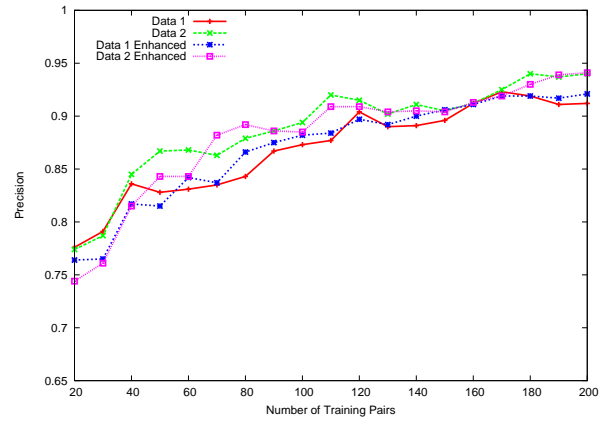


Figure 12: Precision of the enhanced model over the test set with respect to the number of training pairs

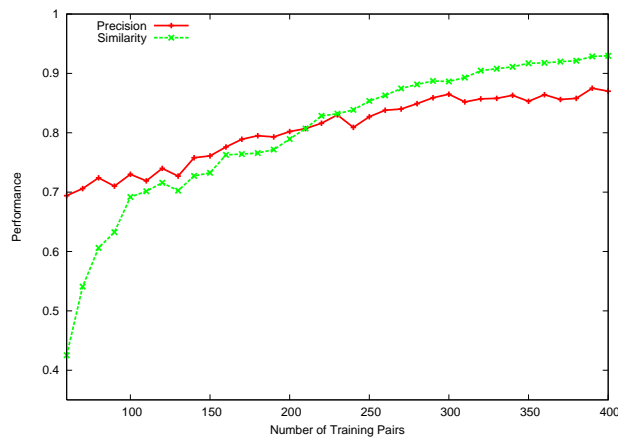


Figure 13: Performance over the test set with respect to the number of training pairs when the grades of documents are not known

of the 22nd international conference on Machine learning, 2005.

- [6] J. J. Louviere, D. A. Hensher, and J. D. Swait. *Stated choice methods: analysis and application*. Cambridge University Press, New York, NY, USA, 2000.
- [7] E. M. Voorhees. Evaluation by highly relevant documents. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, New York, NY, USA, 2001. ACM.
- [8] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th ACM SIGIR*, pages 391–398, New York, NY, USA, 2007.
- [9] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proceedings of ACM SIGIR*, New York, NY, USA, 2007.

- [2] O. Chapelle and Z. Harchaoui. A machine learning approach to conjoint analysis. volume 17, pages 257–264, Cambridge, MA, USA, 2005. MIT Press.
- [3] T. Evgeniou, C. Boussios, and G. Zacharia. Generalized robust conjoint estimation. *Marketing Science*, 24(3):415–429, 2005.
- [4] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.
- [5] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings*